Research Paper | Education | E-ISSN No : 2454-9916 | Volume : 9 | Issue : 6 | June 2023

PREVENTING ALGORITHMIC BIAS: EXPLORING AND ANALYZING EXAMPLES

Christine E. Rowe

Research Scholars Program, Harvard Student Agencies, In collaboration with Learn with Leaders

ABSTRACT

Background: This research paper examines various instances of Algorithmic Bias and their adverse effects on individuals and systems. It investigates the connection between these biases and existing laws while proposing solutions to mitigate algorithmic bias. The paper begins by elucidating the four characteristics of an algorithm, namely finiteness, unambiguity, well-defined inputs and outputs, and feasibility. Furthermore, it provides an explanation of different types of algorithmic bias, including programmed bias, training data bias, interpretation bias, algorithmic focus bias, algorithmic processing bias, and transfer context bias. The study delves into three prominent examples of algorithmic bias within the domains of medicine, law enforcement, and education. Additionally, the paper addresses racial and financial bias concerns, examines relevant privacy laws, and explores strategies for coexisting with algorithms.

KEYWORDS: Algorithmic Bias, Algorithms, Characteristics, Types of Bias

INTRODUCTION

Preventing Algorithmic Bias: Exploring and Analyzing Examples

Algorithms are a fundamental aspect of our world. They run our social media, filter job applicants, and give us ads based on our interests. Most of the time, we are oblivious to them because they do their job. But what about the times when they are unable to play fair and develop biases? How does it happen? Are we responsible? To what extent can algorithmic bias be prevented when the algorithm learns from the data it collects?

In order to understand the impact algorithms have on us, we need to understand them. First, algorithms contain four main characteristics (Diploma Programme, 2022, 2):

1. Finite: an algorithm must have finite steps to complete a task.
2. Unambiguous: an algorithm's instructions must be clear and straightforward.
3. Well-Defined Inputs and Output: an algorithm must have clear data and a clear product.
4. Feasible: an algorithm must be able to function within the realm of logic.

The different characteristics impact how algorithms affect the real world. Algorithmic bias, on the other hand, occurs when an algorithm's outputs become unequal, often favoring certain groups of people, with racial bias being a prominent concern.

To understand the negative impacts of algorithmic bias, it is important to understand the different types of biases possible. These biases can be caused by a multitude of reasons (Danks, D., & London, A. J., 2017, pp. 4691-4697):

1. Programmed Bias: This is when programmers will program their own biases into an algorithm. It can be an intentional bias or a subconscious bias. Either way, the algorithm is compromised because the code has bias itself.
2. Training Data Bias: When an algorithm is created it goes through rounds of testing to 'learn' how to do its job, and it uses pattern recognition to create different outputs. Then it is told if those outputs were correct. When testing humans, this can interfere by choosing selective data. For example, using primarily white people to test a facial recognition algorithm can create an algorithm that struggles to recognize the faces of other races. Selective data makes it difficult for an algorithm to process unrecognizable data, which later leads to an unfair output.
3. Interpretation Bias: This is when the algorithm in use is misunderstood by the user. Algorithms are only effective if the user can correctly interpret the information.
4. Algorithmic Focus Bias: This happens when an algorithm is given training data that does not relate to its purpose. It creates different categories that do not relate to the main goal and can affect the outputs.
5. Algorithmic Processing Bias: This is when an algorithm's product uses statistics to create predictions. These predictions may be based on unreliable information, outdated sources, and opinions. Statistically biased estimators can be used on small scales, but they become unreliable the larger they become.
6. Transfer Context Bias: When an algorithm is used outside its intended context it can become biased. The algorithm itself may be completely ethical but if it is transferred to a new environment, it becomes unreliable.

These multiple reasons for algorithmic bias create various biases that impact different aspects of our lives. Understanding real-world examples helps illustrate how these biases affect individuals.

Examples of Algorithmic Bias

1. *Medical Field*

In a 'high-risk care management' program, an algorithm was employed to assess patients and determine their eligibility for special care. The aim of such programs is to ensure that high-risk patients receive timely and appropriate care, thus increasing treatment accessibility and minimizing costs. The algorithm's task was to identify patients who were most vulnerable and in need of priority care.

To make its decisions, the algorithm relied on input data related to the cost of care. The underlying assumption was that patients requiring more expensive care were in greater need. However, this approach resulted in several problems, particularly a bias related to race. On average, individuals identifying as black generated $1,144 less in healthcare costs than those identifying as white. As a result, the algorithm exhibited bias and allocated less care to individuals identifying as black. It became evident that predicting health outcomes, rather than cost, was a more effective determinant of necessary care (Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S., 2019).

While the algorithm facilitated assistance for many individuals, it also hindered others from receiving the necessary care. Understanding this bias and making appropriate modifications can render the algorithm highly beneficial. For instance, by considering the nature and severity of ailments or collaborating with human professionals, algorithms can serve as tools to improve existing systems and augment the decision-making process.

2. *Law Enforcement*

In collaboration with law enforcement, an algorithm was utilized for suspect identification through facial recognition. The algorithm analyzed security footage and compared recognized faces with IDs within its jurisdiction. Law enforcement officers provided the algorithm with information about the suspect's face, features, and the committed crime. Based on this input, the algorithm generated a list of potential matches and predicted the likelihood of the suspect's involvement based on their criminal record.

In one particular case, a retail store robbery occurred in October 2018, where watches worth $4,000 were stolen. The store's security camera captured the crime. In January 2020, Robert Williams, a 42-year-old man, was arrested for the offense. The algorithm had matched Williams' face from the footage with his driver's license photo, leading to his identification and subsequent arrest. However, there was a significant problem—Williams had a solid alibi at the time of the robbery and bore no resemblance to the person depicted in the video (Fussell, 2020).

This algorithm relied on physical attributes and records to determine the likelihood of a suspect being a match. However, both aspects can be flawed. Algorithms operate in a literal manner, analyzing raw data without considering contextual factors. For instance, facial expressions can influence the appearance of a person's face, and intense emotions may not align with an ID photo. Moreover, having a criminal record does not necessarily indicate an individual's propensity to commit a crime.

The algorithm's fatal error occurred because it did not account for human intervention. By relying solely on the program's output, Williams was wrongfully detained for 30 hours, highlighting the importance of human oversight in algorithmic decision-making processes.

3. *Education*

Proctorio is an algorithm specifically designed for remote test-taking, aiming to monitor students' faces during exams to detect any signs of cheating. It tracks students' facial movements, particularly if they divert their gaze from the exam. However, this algorithm has faced criticism from students like Lucy Satheesan at Miami University in Oxford, Ohio, who perceived it as an invasion of privacy. Satheesan conducted research on Proctorio and discovered that it employed an open-source code called OpenCV, which is a computer vision program utilized for facial recognition.

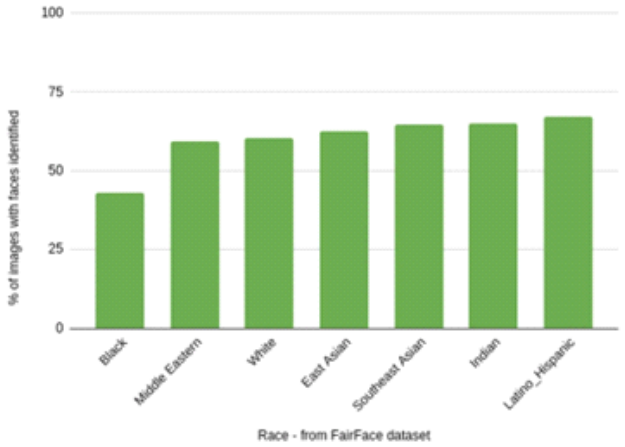

Figure 1: Proctorio Facial Recognition - Detection Rates
Source: Satheesan (2021)

Through her experiments using OpenCV, Satheesan found that Proctorio exhibited limitations in accurately recognizing faces (refer to Figure 1), especially those with darker skin tones. She observed that black faces were recognized inaccurately over 50% of the time, highlighting the algorithm's overall deficiency in facial recognition. On the other hand, the highest recognition rate was for Latino individuals, just under 75% (Satheesan, 2021).

In theory, Proctorio aimed to ensure test integrity for remote learners, providing a more reliable testing environment outside of physical classrooms. However, in practice, the algorithm flagged black students as potential cheaters. This example illustrates how Proctorio, as a facial recognition program, was likely trained and tested primarily on individuals with lighter skin tones. Algorithms rely on pattern recognition based on their training data. If the algorithm lacks diversity in its training data, it becomes unreasonable to expect accurate performance across different racial groups. The limited representation of darker-skinned individuals in the training data led to a bias in favor of students with lighter skin.

Laws Surrounding Data Privacy

1. *GDPR - European Union (EU)*

One significant legislation addressing data privacy is the General Data Protection Regulation (GDPR), which applies within the European Union (EU). This law aims to safeguard individuals' online privacy and has implications for algorithms. According to GDPR, algorithms can only utilize personal data when it is necessary for the intended purpose, and explicit consent from the user must be obtained. The companies responsible for developing and deploying algorithms are obligated to protect the rights and freedoms of their users. Additionally, the EU has specifically stated that algorithms are prohibited from processing information provided by children, as minors are unable to provide informed consent and are considered a more vulnerable group within society (Are There Restrictions on the Use of Automated Decision-Making?, n.d.).

2. *CCPA - California, USA*

Out of the 50 states in the United States, California has emerged as a frontrunner in online privacy regulations with the California Consumer Privacy Act (CCPA). This law grants private citizens certain rights concerning their personal data. It includes provisions allowing individuals to correct inaccurate data about themselves and to disclose and release data to companies and the public. The CCPA emphasizes the right to be informed about any personal data collected, the right to delete collected personal data, the right to opt-out of data sharing, and protection against discrimination for exercising CCPA rights (California Consumer Privacy Act (CCPA) | State of California - Department of Justice - Office of the Attorney General, 2023).

These laws illustrate how governments are striving to provide safeguards for their citizens against potential risks associated with technologies. However, there is still much to be understood by governments in addressing all potential biases. Examining real-world examples and their connections to biases can contribute to the development of more comprehensive and protective legislation.

Connecting to Types of Algorithmic Bias

1. *Medical - Algorithmic Processing Bias*

The algorithm employed in the context of 'high-risk care management' can be classified as Algorithmic Processing Bias. This algorithm relied on financial predictions to establish a system where individuals with a higher likelihood of spending more money would receive greater assistance. Consequently, the predicted financial factors had a direct influence on the allocation of resources to those in need. Notably, the algorithm exhibited a bias towards prioritizing individuals of white ethnicity due to the average higher healthcare expenditures associated with this demographic.

2. *Law Enforcement - Algorithmic Processing Bias*

The algorithm employed in Law Enforcement can be classified as Algorithmic Processing Bias. Robert Williams was directly affected by the algorithm's inability to accurately process the data in accordance with the circumstances. Specifically, the facial recognition system failed to correctly analyze the security video, leading to an erroneous prediction that Robert Williams was the individual depicted in the footage.

3. *Education - Training Data Bias*

The algorithm employed in Education can be categorized as Training Data Bias. During Lucy Satheesan's testing of OpenCV, it was observed that the program exhibited limitations in accurately processing faces. Notably, it struggled to process the faces of individuals with darker skin tones. This inability to properly process faces indicates that the source code was tested primarily on individuals with fair complexions.

Solution

While there are numerous examples of algorithmic bias negatively impacting individuals, focusing solely on the problem overlooks the opportunity to enhance algorithms for greater equity. Algorithms are deeply ingrained in our society, necessitating a path forward that prioritizes fairness. Preventing algorithmic bias entails treating algorithms as tools rather than definitive solutions. Algorithms operate in a concrete and literal manner, processing data without the ability to comprehend nuance. On the other hand, humans possess empathy and can understand the complexities of a situation. When algorithms are given decision-making authority, they manifest the negative aspects of human subjectivity, amplifying potential biases. Harm arises when algorithms are solely relied upon for decision-making. Algorithms simply execute the instructions they are programmed with and lack consciousness or nuanced understanding.

However, when algorithms are utilized as tools, rather than sole decision-makers, it enables human capacity for nuance to prevail. This approach would have made a significant difference for individuals like Robert Williams if human reviewers had examined the footage and considered additional factors. Such an assessment would have clearly demonstrated the lack of an accurate match, sparing Williams the fear and frustration of being held for an unnecessary 30 hours. Viewing algorithms as tools allows for automated assistance while placing decision-making power in the hands of humans, who can weigh data with empathy and nuance.

Conclusion

Algorithms play a significant role in our world, and their blind use in decision-making processes sets the stage for algorithmic bias. This bias often leads to a range of problematic and inequitable outcomes, disproportionately affecting minority groups and the more vulnerable segments of society. However, by recognizing algorithms as tools and acknowledging their potential biases, while retaining decision-making authority in the hands of humans who possess the capacity for nuance and empathy, we can mitigate some of these biases. Although much remains to be understood and accomplished in the realm of algorithms, through critical thinking and a commitment to responsible and equitable use, we can continue to learn and adapt. Furthermore, fostering a deeper understanding of algorithmic bias and enacting protective legislation will be crucial steps toward a more just and inclusive algorithmic future.

REFERENCES

1. Are there restrictions on the use of automated decision-making? (n.d.). European Commission. Retrieved March 19, 2023, from https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/dealing-citizens/are-there-restrictions-use-automated-decision-making_en
2. California Consumer Privacy Act (CCPA) | State of California - Department of Justice - Office of the Attorney General. (2023, February 15). California Department of Justice.

Retrieved March 19, 2023, from https://oag.ca.gov/privacy/ccpa
3. Clark, M. (2021, April 8). Students of color are getting flagged to their teachers because testing software can't see them. The Verge. Retrieved March 18, 2023, from https://www.theverge.com/2021/4/8/22374386/proctorio-racial-bias-issues-opencv-facial-detection-schools-tests-remote-learning
4. Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In Ijcai (Vol. 17, No. 2017, pp. 4691-4697)
5. Diploma Programme. (2022). Digital Society Guide. International Baccalaureate Organization. Fussell, S. (2020, June 24). A Flawed Facial-Recognition System Sent This Man to Jail. WIRED. Retrieved March 18, 2023, from https://www.wired.com/story/flawed-facial-recognition-system-sent-man-jail/
6. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
7. Satheesan, L. (2021, March 18). Proctorio's facial recognition is racist. — Proctor Ninja. Proctor Ninja. Retrieved March 19, 2023, from https://proctor.ninja/proctorios-facial-recognition-is-racist